DOCUMENT RESUME

ED 287 284 FL 016 940

AUTHOR Madsen, Harold S.
TITLE Utilizing Rasch Analysis to Detect Cheating on Language Examinations.
PUB DATE 87
NOTE 14p.; In: Language Testing Research; Selected Papers from the Colloquium (Monterey, California, February 27-28, 1986); see FL 016 938.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cheating; English (Second Language); Evaluation Methods; *Language Tests; *Multiple Choice Tests; *Probability; Responses; Second Language Instruction; Statistical Analysis; Testing Problems; *Test Interpretation; *Test Use
IDENTIFIERS *Rasch Model

ABSTRACT

A study investigated the effectiveness of the Rasch procedure in measuring response appropriateness, especially for the detection of cheating on multiple-choice language tests. The report gives background information on appropriateness measurement and its potential uses, reviews recent research on cheating and its detection, and describes three contrived studies, one involving English as a second language (ESL), to determine the Rasch model's usefulness. While the Rasch one-parameter model can detect some aberrations on test response due to cheating, the study of ESL cheating highlights its limitations in pinpointing cheating as contrasted with other causes of inappropriate responses. Further research into both evaluation inequities and cheating is recommended. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED287284

UTILIZING RASCH ANALYSIS TO DETECT CHEATING ON LANGUAGE EXAMINATIONS

Harold S. Madsen
Brigham Young University

The purpose of this investigation is to assess the effectiveness of the Rasch procedure in conducting appropriateness measurement--notably the detection of cheating on multiple-choice language tests. Of the several statistical models that have been developed to further latent trait analysis related to objective tests and survey-instruments, the Rasch procedure seems to be the most appropriate option for examining the small data samples (say 50 to 100 subjects) often encountered in ESL and FL settings (Lord 1983, Henning 1984).

Unlike multiple-parameter statistical models which incorporate factors such as discrimination and guessing into their formulas, the Rasch one-parameter model is based on the single factor of response probability--the likelihood of each individual's answering a given item correctly and the corollary probability of items being responded to correctly. With the logarithmic calculations built into the Rasch model, it is possible to generate 1) remarkably precise calibrations of item difficulty and examinee performance, 2) item calibrations essentially independent of the sample populations being measured, and 3) measures of aberrant or unexpected behavior, such as the tendency for a difficult item to be answered correctly by some low-proficiency examinees.

Rasch analysis has been applied to a wide variety of testing needs, including item banking (Woods and Baker 1985, Henning 1986), sample-free item analysis (Wright and Masters 1982), and identification of item bias (Hulin, Drasgow and Parsons 1983, 167-170; Madsen and Larson 1986). It has even been used to calibrate speaking test interview items allowing for partially-correct responses (Griffin 1985) and to analyze a scalar self-rating language ability instrument and a writing scale (Davidson and Henning 1985, Henning and Davidson 1986).

Appropriateness Measurement

Another potential Rasch application of considerable interest is that of appropriateness measurement, a broad area of investigation only part of which is concerned with the problem of cheating. Two important recent texts each devote a major chapter to appropriateness measurement--Item Response Theory: Application to Psychological Measurement (Hulin and others 1983, 110-151) and New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing (Weiss 1983, 109-131). The former authors define such research as "any method of detecting individuals for whom test scores fail" as measures of the latent trait, ability, or interest (1983, 111). Investigations in this area examine all sorts of test taking anomalies. These aberrations can be roughly classified into two categories: unintentional behaviors and deliberate tampering with test responses.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Schiffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL016940

Relative to the first category, Wright and Stone (1979, 165-180), Levine and Drasgow (1982), Hulin and others (1983, 112-114) mention several examples of unintentional behaviors that can result in misleading test scores. These include a student's guessing, being inattentive or bored, "fumbling" because of confusion over the intent of test questions or exam directions, stumbling on items because of their cultural bias, not fully completing an exam, mismarking a machine-scorable answer sheet, or simply looking at items (as on an IQ instrument) from a novel or creative point of view. Since he is concerned with the problem of inappropriate score results, Wainer challenges the usual psychometric procedure of point reduction for coping with guessing. He posits that behavior labeled as guessing or "sleeping" (carelessness due to inattentiveness or boredom) could in fact be a result of "unusual patterns of knowledge" or other factors (1983, 64-65). In their pioneering article on appropriateness measurement, Levine and Rubin refer to those with deficient test taking strategies: i.e., individuals inclined toward "excessive guessing, reluctance to answer questions unless absolutely sure, [and] ... behavior suggesting lack of experience with test format" (1979, 270).

Unintentional aberrant responses can also stem from special handicaps: for example, a person with limited language proficiency taking a test designed for native speakers (Oller and Perkins 1978, Levine and Rubin 1979), or a student with hearing deficiency taking an aural comprehension test. In addition, a person with an atypical educational background may be handicapped on a test designed for mainstream students (Drasgow 1982).

Turning to the second category of appropriateness measurement, we note that deliberate altering of results also takes many forms. Students recently have tampered with computer files of tests: a graduate student programmed a computer to give himself right answers and other students incorrect answers; in Queens, New York, students used computers to raise 10,000 grades (Moody 1980), and teachers have even tampered with student answer sheets, filling in correct responses for deleted items (Jensen 1980). There are numerous occasions on which answer sheets have been obtained and sold to some candidates in advance of the test. Examinees have also been known to suppress scores on a minimal proficiency examination such as those administered by the government for military draftees (Olsen 1986). More common is an examinee's copying of a number of items from another student during a test. Responses on attitude measures are also at times modified by the respondent in order to match perceived expectations on the part of those using the instrument.

Recent Research on Cheating

A variety of recent studies have disclosed that cheating is a serious problem in today's schools (Roizen 1982, Singhal and Johnson 1983, Nuss 1984). Others have sought to determine the causes of dishonest behavior in the classroom (Shab 1980, Chapman 1980), and even the personality traits of cheaters (Leming 1980, Antion and Michael 1983, Forsyth, Pope and McMillan 1985). In addition, possible solutions to the cheating problem have received considerable attention (Houston 1977, Urbain 1980, Guttman 1984). A related research concern is that of cheating <u>detection</u>. The following studies focus on this effort.

Utilizing IRT Procedures to Detect Cheating

A variety of conventional procedures for identifying cheating have been employed over the years. One statistically-based procedure utilized to verify suspected cheating focuses on the match of wrong answers between the suspected cheater and his or her "source" plus the wrong-answer patterns of the remaining students (Cody 1985, 136-137). A variation on the procedure utilizes students in adjacent seats as sources and two random papers three or more seats away for the comparison (Houston 1983).

Hulin and others after surveying a variety of traditional or heuristic approaches designed to identify inappropriate test performance (Ghiselli's prediction of predictability, Jacob's weighted average, as well as Donlon and Fischer's personal biserial) find them imprecise and largely ineffective in identifying normal and aberrant behavior patterns (1983, 114-121).

On the other hand, recent IRT research in the area of appropriateness measurement has yielded very impressive results in identifying both normal response patterns and aberrant patterns. Forerunners in this movement, such as Levine and Rubin, repeatedly assure readers that they are not committed to the indices they have developed (1979, 286). But there tends to be some reservation among these scholars about the effectiveness of the one-parameter model: For example, in his dissertation research, Drasgow found the three-parameter model superior to the one-parameter model in distinguishing between normal and aberrant response patterns on GRE multiple-choice Verbal Aptitude items (1978).

But Drasgow does stop short of advocating the exclusive use of a three-parameter model, particularly if cost is a factor (1982, 305-307). Moreover, Hulin and others indicate that differences between the three- and one-parameter models are not sufficient to warrant the use of the former model "under all circumstances" (1983, 141). Given the suitability of the one-parameter model for small-sample test studies, it was decided to undertake our appropriateness research by utilizing the Rasch model.

Studies on Cheating

Studies Involving TOEFL Cheaters

Since cheating on exams is such a widespread problem, it was enticing to determine whether principles of appropriateness measurement could be utilized as a kind of "radar" to detect such aberrant performance. At the outset, however, it was obvious that our "radar" could not be expected to detect all dishonest performance on a language exam: While it should be possible to detect a weak student's copying a page or two from a much superior student, it wouldn't be possible to detect an advanced language learner copying from another advanced student (Levine and Rubin 1979, 270), because their pattern of expected responses would be very similar. Moreover, if Student A took an exam on behalf of Student B, it would be possible to detect this only if Student A performed abnormally, for example by arbitrarily suppressing his score.

Being aware that Educational Testing Service had recently apprehended some cheaters on the TOEFL--individuals who had received payment to take the exam for foreign candidates seeking admission to American universities--the author arranged to attempt their identification by means of Rasch analysis. It was hypothesized that to escape detection these professional cheaters would suppress their scores. Arbitrarily missed items would then flag the cheater as an outlier on the computer printout--an able student missing a number of relatively easy items. While the initial investigations failed to identify cheaters, they did set the stage for follow-up work which was successful in spotting the aberrant behavior of persons tampering with test items.

TOEFL Study One. The first study involved two known cheaters on the TOEFL (Test of English as a Foreign Language). Two sets of answer sheets, with identification removed, were obtained from Educational Testing Service, the agency responsible for preparing and administering TOEFL. Selected from exam papers at five centers, there were 29 test papers from TOEFL Form 7; and there were 30 from TOEFL Form 15. Each of the two sets of answer sheets included the paper of one unidentified cheater.

The two groups were Rasch analyzed separately utilizing the Microscale Plus routine developed by Wright and Linacre (1985) for the IBM personal computer. The outlier-sensitive fit statistic (or "outfit") was obtained, since this identifies unexpected events, such as examinees performing in an aberrant fashion. But contrary to expectations, no clearly identifiable outliers were identified (Figure 1). In short, the cheaters had not been detected. The reason for this became apparent, however, after their ID numbers were provided by ETS: They had performed so well (with scores of 610 and 623) that it was apparent they could not have suppressed their scores.

TOEFL Study Two. ETS next agreed to send two additional groups of papers (identification removed as before), with cheaters whose scores were well under 600: one with a score of 557 and the other with a score of 563. The paper of one cheater was included in a group of 46 answer sheets from TOEFL Form 8; the paper of the second cheater was included in a batch of 44 answer sheets from TOEFL Form 9. As in the first TOEFL study, data from five different testing centers were utilized.

Again the outfit plots (Figure 2) generated by Rasch analysis of these second two groups failed to identify the two known cheaters. Interestingly, however, the analysis did identify an outlier with a dramatically high outfit score. For such a surprising outfit measure, this weak student (number 13) would need to exceed the chance probability of answering correctly a number of items well beyond his level of ability. The most plausible explanation was that he had cheated. But his low score informs us that this would not have been the result of score suppression (the object of this investigation). This probable cheater was reported to TOEFL officials.

Our initial investigation, then, into the response patterns of known cheaters reveals that either they had not attempted to suppress their scores, or that such an effort had been so minimal that the analysis was unable to detect it.

Cheating Studies Utilizing Artificial Data

Contrived Cheater Study One. This investigation attempted to determine how effective Rasch analysis would be in identifying a small percentage of manipulated scores. Drawing upon data from 183 ESL students, the author selected two papers at the low end of the scale and rescored them so that correct answers were given to five items in the top quartile of difficulty (see Madsen and Larson 1986, 59). When the data were Rasch analyzed again, the outfit scores for these two papers had jumped dramatically: from approximately 1 to well over 4 -- the highest outliers of the entire group of 183 students (Figure 3).

Contrived Cheater Study Two. In contrast with this spuriously high score situation, it was next decided to devise a more realistic setting, one parallel to that hypothesized for professional cheaters on the TOEFL. The cheater's paper from the TOEFL Form 7 group was selected, and a "miss" of one in every ten responses was manipulated. (These incorrect responses plus those missed as a matter of course by this non-native cheater resulted in a TOEFL score of 577). The data were again Rasch analyzed. Now with a suppressed score, the paper had the highest outfit of any in the group (Figure 4).

But while such an outfit score could help corroborate allegations of cheating, it might not always flag decidedly aberrant test behavior. Like Levine and Rubin (1979), we needed to evaluate a situation well beyond the 10 percent level of manipulation. After examining 4, 10, 20 and 40 percent score modifications of artificial data, they had found that only when tampering reached the 20 percent level or higher did they achieve good detection rates.

Contrived Cheater Study Three. Now a TOEFL scenario was devised, involving a supposed native speaker cheater with a 20 percent score suppression. This contrived test paper was added as examinee number 30 in the TOEFL Form 7 group. The data were again Rasch analyzed, this time with impressive results. The native-language "cheater" had an outfit score approximately three times as great as the highest regular student (Figure 5).

Discussion

The appropriateness-measurement investigations reported here demonstrate that the Rasch one-parameter model is able to detect some aberrations in test responses due to cheating. The TOEFL studies remind us, however, of the limitations of IRT "radar": Unless spuriously high or spuriously low scores are generated, the cheater will go undetected. The studies involving score manipulation point out a second limitation: Unless the score suppression or inflated scores due to copying or memorizing a block of answers are substantially above the 10 percent level, there is a danger of "false arrest." But as Levine and Rubin found with another model (1979), when the alterations approach the 20 percent level or beyond, Rasch analysis can easily detect aberrant performance. Caution is needed, of course, in attributing all extreme outlier or aberrant performance to cheating. As we have seen, inappropriate performance on a test could result from a variety of causes.

The door is now open for testing organizations, educational institutions, and even a great many individual teachers to deal more thoroughly and

systematically than ever before with faulty measurement on language exams. Not only are we more able to apprehend cheaters but through IRT procedures such as Rasch analysis we can help rectify inequities in the evaluation process. As valuable as the new technology is, however, it will need to be accompanied by human investigation into the causes of the aberrant performance.

References

Antion, David L. and William B. Michael; 1983. Short-term predictive validity of demographic, affective, personal, and cognitive variables in relation to two criterion measures of cheating behaviors. Educational and Psychological Measurement 43(2): 467-482.

Chapman, William; 1980. A cheating problem in Japan. Today's Education 69(4): 42.

Cody, Ronald P; 1985. Statistical analysis of examinations to detect cheating. Journal of Medical Education 60(2): 136-137.

Davidson, Fred and Grant Henning; 1985. A self-rating scale of English difficulty: Rasch Scalar Analysis. Language Testing 2:2 (December), 164-179.

Drasgow, Fritz; 1978. Statistical Indices of the appropriateness of aptitude test scores. Unpublished Ph.D. dissertation, University of Illinois at Urbana-Champaign.

Drasgow, Fritz; 1982. Choice of test model for appropriateness measurement. Applied Psychological Measurement 6:3 (Summer), 297-308.

Forsyth, Donelson R., William R. Pope, and James H. McMillan; 1985. Students' reactions after cheating: an attributional analysis. Contemporary Educational Psychology (1): 72-82.

Griffin, Patrick E.; 1985. The use of Latent Trait Models in the calibration of tests of spoken language in large-scale selection-placement programs, 149-161. New Directions in Language Testing. Y. P. Lee, Angela C.Y.Y. Fok, Robert Lord and Graham Low (Eds.) Oxford: Pergamon Press.

Guttman, Joseph; 1984. Cognitive morality and cheating behavior in religious and secular school children. The Journal of Educational Research 77(4): 249-254.

Henning, Grant; 1984. Advantages of Latent Trait Measurement in language testing. Language Testing 1(2): 123-133.

Henning, Grant; 1986. Item banking via DBASE II: The UCLA ESL Proficiency Examination experience. Technology and Language Testing. Charles W. Stansfield (Ed.) Washington, D.C.: TESOL.

Henning, Grant and Fred Davidson; 1986. Scalar analysis of composition ratings. Paper presented at the Eighth Language Testing Research Colloquium, Monterey, Ca.; February 28 (This volume).

Houston, John P; 1977. Cheating behavior, anticipated success-failure, confidence, and test importance. Journal of Educational Psychology 69(1): 55-60.

Houston, John P.; 1983. Alternate test forms as a means of reducing multiple-choice answer copying in the classroom. Journal of Educational Psychology 75(4): 572-575.

Hulin, Charles L., Fritz Drasgow, Charles K. Parsons; 1983. Item Response Theory: Applications to Psychological Measurement. The Dorsey Professional Series. Homewood, Illinois: Dow Jones-Irwin.

Jensen, A.R.; 1980. Bias in Mental Testing. New York: Free Press. Reported in Hulin, Charles L., Fritz Drasgow, and Charles K. Parsons; 1983. Item Response theory: Application to Psychological Measurement. Homewood, Illinois: Dow Jones-Irwin.

Leming, James S.; 1980. Cheating behavior, subject variables, and components of the internal-external scale under high and low risk conditions. Journal of Educational Research 74(2): 83-87.

Levine, Michael V. and Fritz Drasgow; 1982. Appropriateness measurement: review, critique and validating studies. British Journal of Mathematical and Statistical Psychology 35:42-56.

Levine, Michael V. and D.F. Rubin; 1979. Measuring the appropriateness of multiple-choice test scores. Journal of Educational Statistics. 4(4): 269-290.

Lord, Frederic M.; 1983. Small N justifies Rasch Model, 51-61. New Horizons in Testing: Latent Trait Theory and Computerized Adaptive Testing, David J. Weiss (Ed.) New York: Academic Press.

Madsen, Harold S. and Jerry W. Larson; 1986. Computerized Rasch Analysis of item bias in ESL tests, 47-67. Technology and Language Testing, Charles W. Stansfield (Ed.) Washington, D.C.: TESOL

Moody, Sid; 1980. Cheating: a national pastime. Today's Education 69(4): 32.

Nuss, Elizabeth M.; 1984. Academic integrity: comparing faculty and student attitudes. Improving College and University Teaching 32(3): 140-144.

Oller, John W., Jr. and Kyle Perkins; 1978. Language in Education: Testing the Tests. Rowley, Massachusetts: Newbury House Publishers, Inc.

Olsen, James. Personal communication, 1986.

Roizen, M.A.; 1982. The population explosion: test-taking strategies project. School of Education, Hebrew University of Jerusalem, course paper. Reported in Cohen, Andrew D. On Taking Language Tests: What the Students Report. Language Testing 1(1): 70-81.

Schab, Fred. 1980. Cheating in high school: differences between the sexes (revisited). Adolescence 15(Winter): 959-965.

Singhal, Avinash C. and Patti Johnson; 1983. How to halt student dishonesty. College Student Journal 17(1): 13-19.

Urbain, Kermit; 1980. Trust them not to cheat. Today's Education 69 (4):35

Wainer, Howard; 1983. Are we correcting for guessing in the wrong direction? 63-80. New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing, David J. Weiss (Ed.) New York: Academic Press.

Weiss, David J. (Ed.); 1983. New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing. New York: Academic Press.

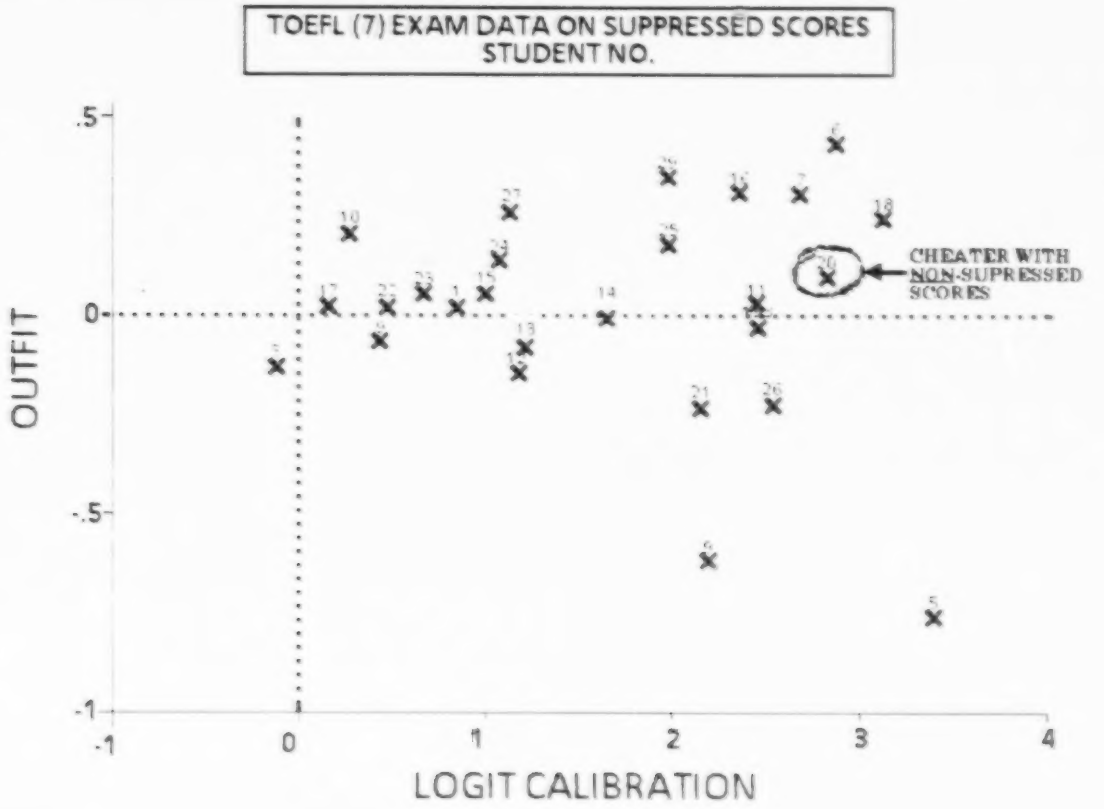
Woods, Anthony and Rosemary Baker; 1985. Item Response Theory. Language Testing 2:2 (December), 119-140.

Wright, Benjamin D. and Michael Linacre; 1985. Microscale Manual. Westport, Ct.: Mediax Interactive Technologies.

Wright, Benjamin D. and Geofferey N. Masters; 1982. Rating Scale Analysis: Rasch Measurement. Chicago: Mesa Press.

Wright, Benjamin D. and Mark H. Stone; 1979. Best Test Design. Chicago: Mesa Press.

Figure 1


TOEFL (7) EXAM DATA ON SUPPRESSED SCORES
STUDENT NO.
CHEATER WITH NON-SUPRESSED SCORES
OUTFIT
LOGIT CALIBRATION
.5
0
-.5
-1
-1
0
1
2
3
4

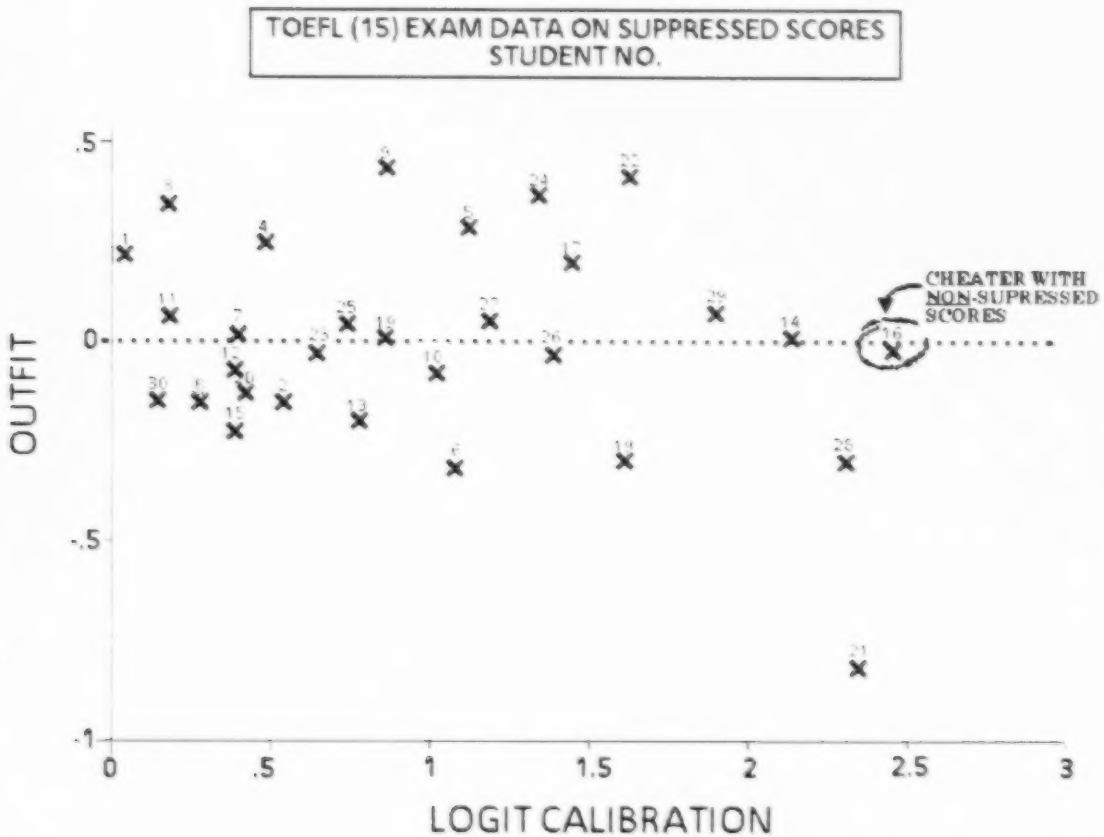
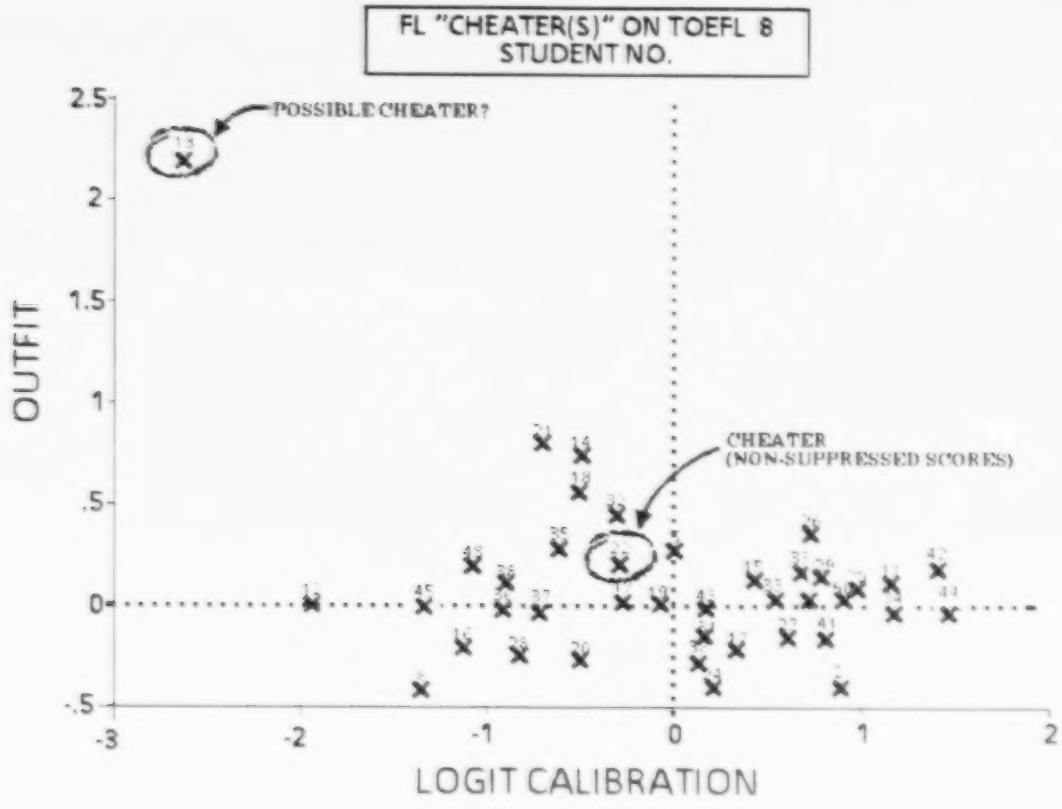

TOEFL (15) EXAM DATA ON SUPPRESSED SCORES
STUDENT NO.
CHEATER WITH NON-SUPRESSED SCORES
OUTFIT
LOGIT CALIBRATION
.5
0
-.5
-1
0
.5
1
1.5
2
2.5
3

Figure 2


FL "CHEATER(S)" ON TOEFL 8
STUDENT NO.
POSSIBLE CHEATER?
CHEATER
(NON-SUPPRESSED SCORES)
OUTFIT
LOGIT CALIBRATION
2.5
2
1.5
1
.5
0
-.5
-3
-2
-1
0
1
2

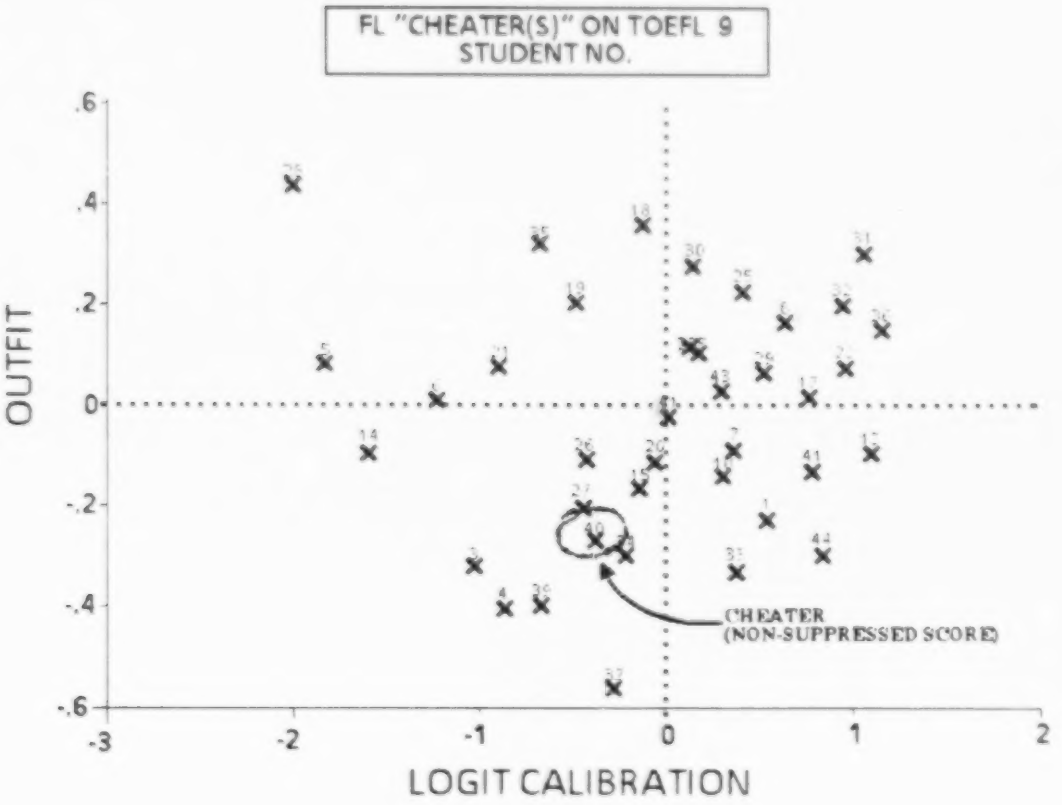

FL "CHEATER(S)" ON TOEFL 9
STUDENT NO.
CHEATER
(NON-SUPPRESSED SCORE)
OUTFIT
LOGIT CALIBRATION
.6
.4
.2
0
-.2
-.4
-.6
-3
-2
-1
0
1
2

Figure 3

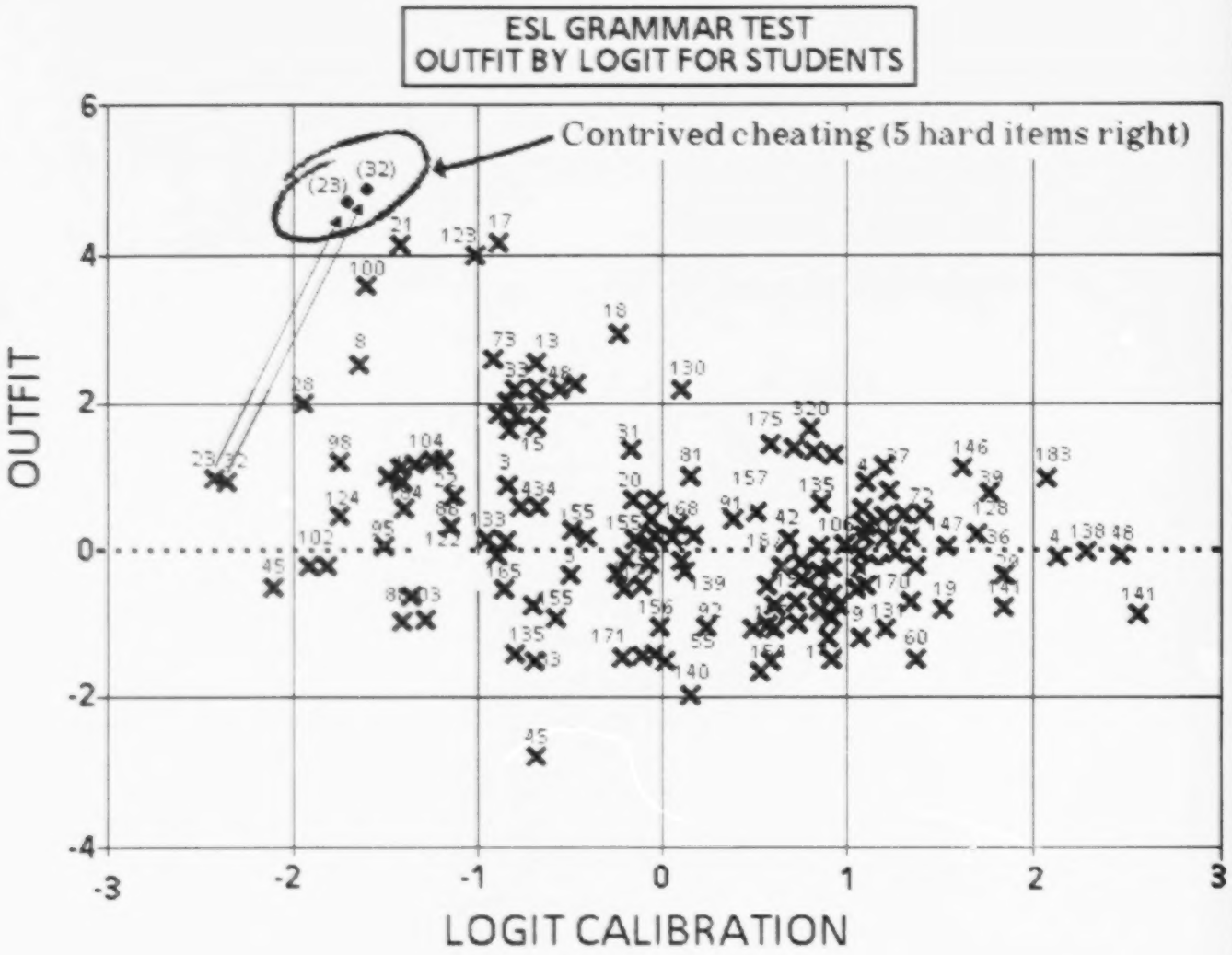

ESL GRAMMAR TEST
OUTFIT BY LOGIT FOR STUDENTS
Contrived cheating (5 hard items right)
OUTFIT
LOGIT CALIBRATION
6
4
2
0
-2
-4
-3
-2
-1
0
1
2
3

Figure 4

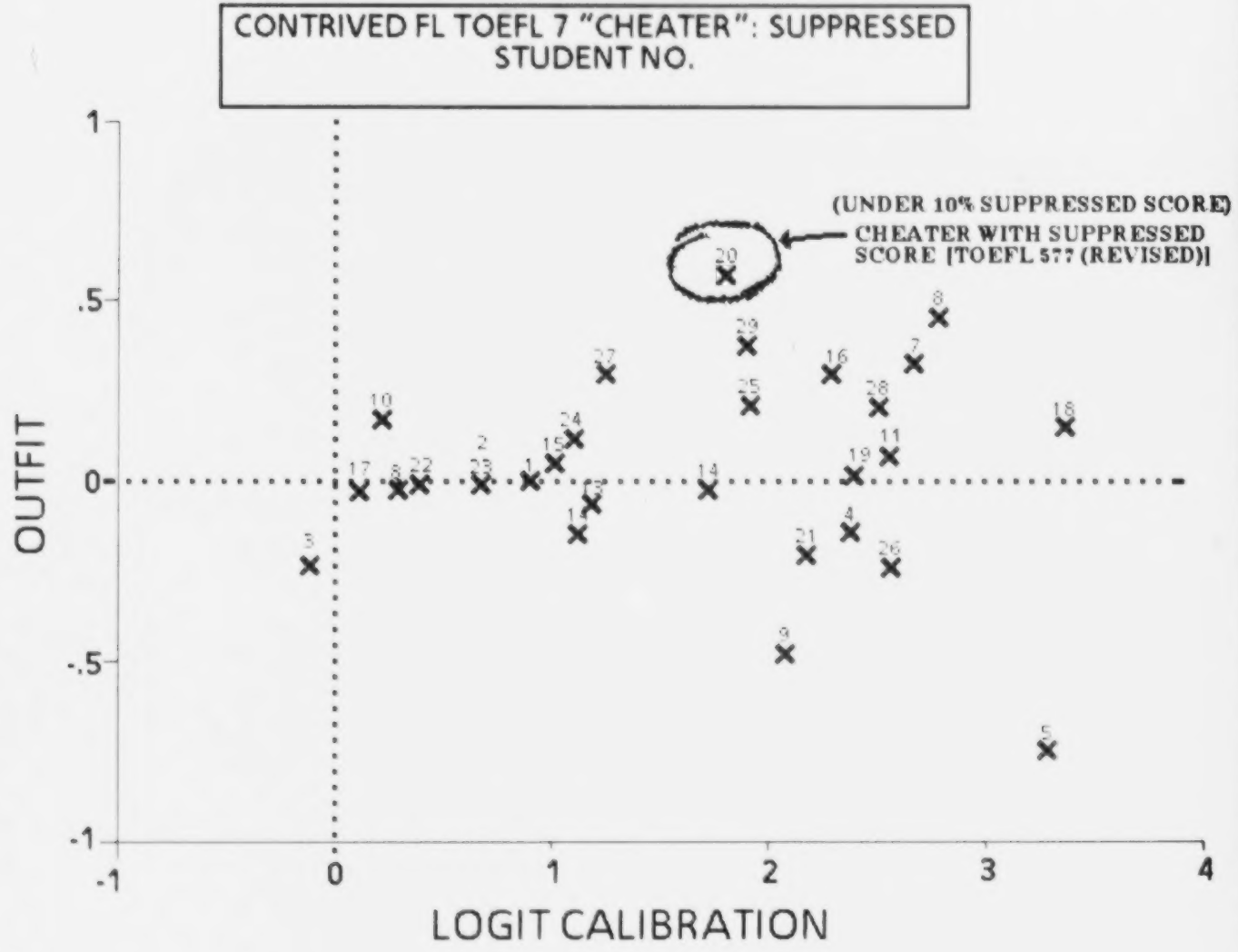
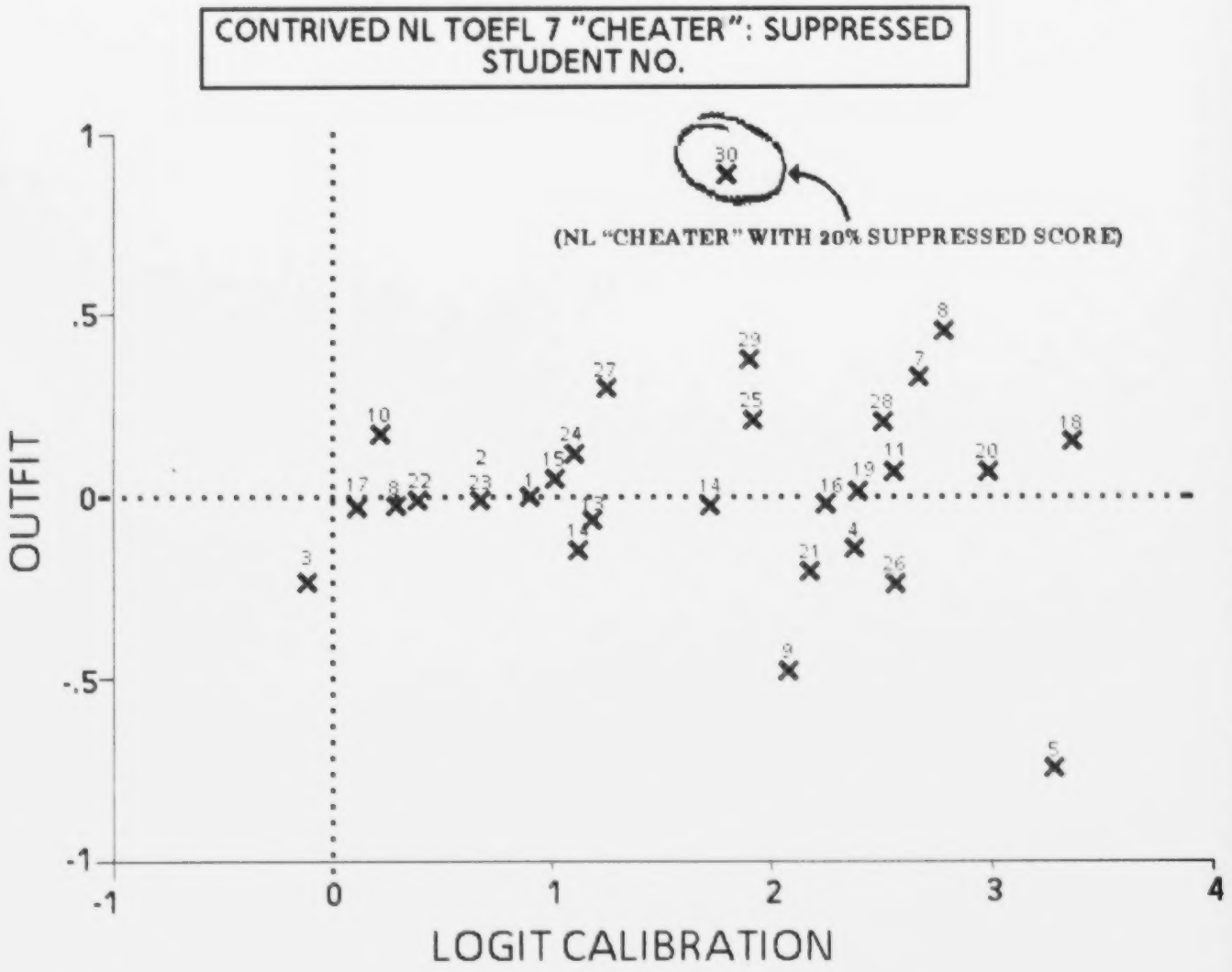

CONTRIVED FL TOEFL 7 "CHEATER": SUPPRESSED STUDENT NO.
(UNDER 10% SUPPRESSED SCORE)
CHEATER WITH SUPPRESSED SCORE [TOEFL 577 (REVISED)]
OUTFIT
LOGIT CALIBRATION
1
.5
0
-.5
-1
-1
0
1
2
3
4

Figure 5


CONTRIVED NL TOEFL 7 "CHEATER": SUPPRESSED
STUDENT NO.
(NL "CHEATER" WITH 20% SUPPRESSED SCORE)
OUTFIT
LOGIT CALIBRATION
1
.5
0
-.5
-1
-1
0
1
2
3
4